Lecture 17 - Multiple Data Sets: Partitions and Mixtures

I. Introduction - As the phylogenomics era unfolds, it's become commonplace to have data from multiple genes.

Even before genomics, it was common situation where one had several sources of data. This may include:

Sequence data from several different genes.
Molecular and morphological data sets.
Behavioral data, molecular data &/or morphological data.

How we should use diverse data in phylogeny estimation has been controversial. There was a tremendously active and (not surprisingly) vehement debate over the best ways to proceed in deriving a phylogeny that incorporates all existing data.

II. Combined vs. Separate Analyses. An issue that received perhaps the most historical attention is whether to take a total evidence approach. Opinions have been strongly divided, at least as far back as the early 90's (there was a symposium at the 1994 Evolution Meetings).

So, let's assume for a minute that we have data from two different genes for a set of taxa.

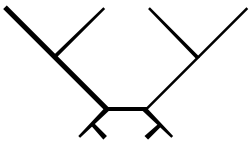
The obvious question that arises is whether we should concatenate those two data sets into a single data set.

There are three historic and one emerging answer to this question:

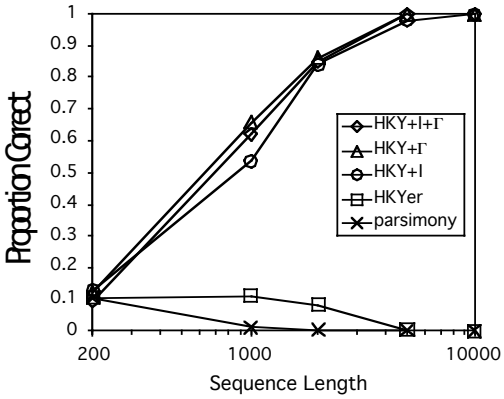
a. "Of course, what an absurd question."
b. "No, we should analyze each separately."
c. "We should do both."
d. Use a multispecies coalescent approach (the topic of the next lecture).

A. Total Evidence – One of the most strongly held positions is that data should always be combined, and the analysis based on all of the characters is always to be preferred.

On the surface, this could be viewed as a simple extension of the idea that more data are better than fewer data.



(a)



(b)

If we think back to our discussion of performance, it may take very many characters (i.e., long sequences) to have a high probability of inferring the true tree.

The idea of combining data from multiple genes makes intuitive sense, given that most genes are far shorter than the 8 –10 Kb length that can be required to have a high degree of confidence in our estimates (at least for short internal branches).

Other (more reasonable) arguments favoring a total evidence approach focus on additivity of signal and hidden support (Gatesy et al., 1999. Cladistics, 15: 21).

The idea here is that there may be signal for particular relationships that is too weak to be detected in analysis of a single gene that nevertheless contributes to support for that group when combined in a simultaneous analysis.

B. Separate Analyses - A few systematists have historically argued against combining data. This is called the **Taxonomic Congruence Approach**.

This is primarily derived from the idea that for molecular data, separate data sets are usually from different genes.

Miyamoto and Fitch (1995. Syst. Biol., 44:64) argue that different data sets are more likely to consist of characters that are mutually independent.

They discuss two reasons that separate data sets may disagree.

1. One (or more) is subject to systematic error. They argue that systematic error should be restricted to a single data set and should not persist across data sets. This may be the result of LBA, or perhaps selection. Thus, separate genes represent natural **process partitions**.

2. Different genes may actually have a different history. We'll discuss the potential causes for this later, but if different genes have different histories, it makes little sense to combine them and thus force them to fit a single bifurcating tree. Only separate analyses (or analyses that permit cycles in the graph) will allow one to detect conflicting signal.

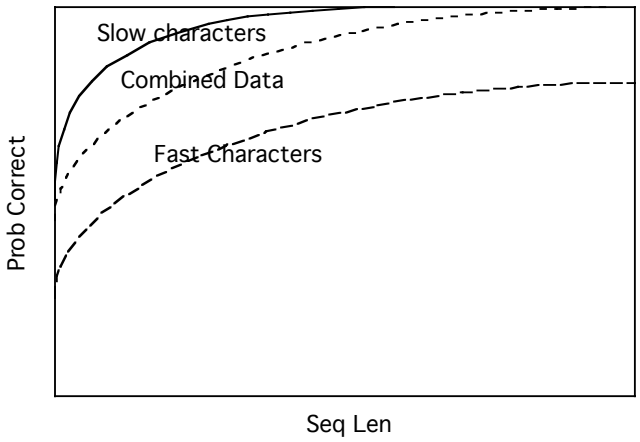
   This can be caused by well-known processes such as incomplete lineage sorting (which we'll address later) and introgression/horizontal gene transfer.

If either of these phenomena occurs, a combined analysis will result in a poor estimate or a wildly inaccurate one.

C. Conditional Combination. The intermediate position is the one with which most phylogeneticists agree. It's sometime called the **Prior Agreement Approach**.

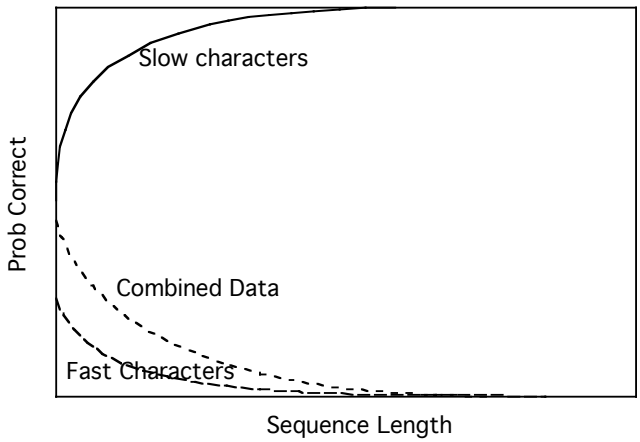
This approach really owes its wide appeal to Bull et al., (1993. Syst. Biol. 42:384). They used simulations to demonstrate that the impact of combining data under several conditions.

For example, if one data set has evolved under conditions where parsimony is both consistent and efficient in phylogeny estimation and another has evolved under conditions that lead to consistent but inefficient estimation, combining data can leads to intermediate efficiency.



This is true even if the two data sets have the same history.

When they simulated one gene under conditions where parsimony is inconsistent and another gene where it is consistent (i.e., the boundary of the Felsenstein Zone), the inconsistent characters can overwhelm the consistent characters.



This has led to the recognition that there may well be situations in which combining data can do more harm than good. Bull et al. used the term **Process Partitions** to describe this scenario.

My take is that Bull et al. (1993) were right that we shouldn't just blindly accept the total evidence tree as the best estimate of the true tree. We should always do separate analyses and combined analyses and evaluate how support for different groupings differs in these.

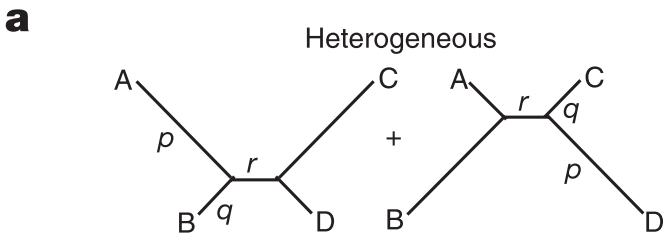
III. Partitioned Models & Concatenated Data

Perhaps because the attitude that more data are necessarily better than fewer data has long been present among phylogenticists, the most common way to treat multilocus data is to concatenate them into a single alignment.

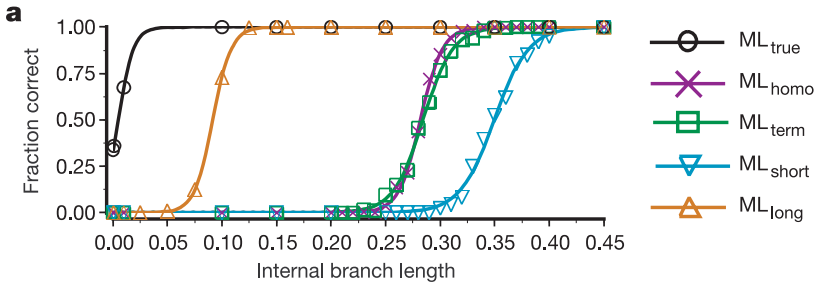
Further, perhaps because of the results like those of Bull et al. (1993), most folks doing so under a model-based perspective (at least ML or Bayesian perspectives) try to accommodate different processes in different genes. We saw this a bit when we discussed the SSR models.

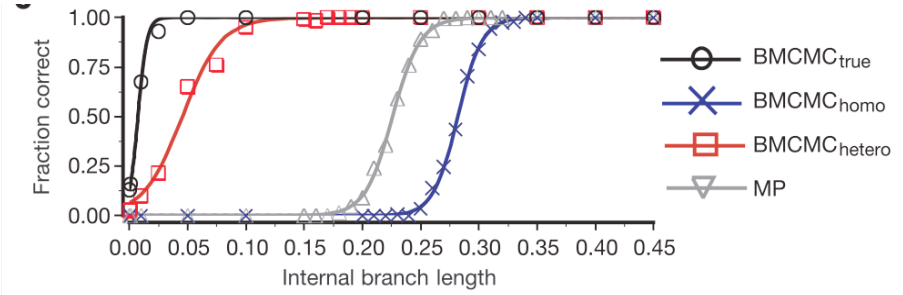
This makes a lot of intuitive sense, because different genes may be under very different evolutionary pressures (i.e. stabilizing versus divergent versus sexual selection), so branch lengths and substitution model may vary considerably from gene to gene.

A. Influence of branch-length heterogeneity was assessed by a really poorly presented paper (Koloczkowski & Thornton, 2004. Nature, 431:980). These authors simulated data partitioned on two sets branch lengths of the same (F-Zone) tree, and with increasing internal branch length. Half the data were simulated on one set of branch lengths and the other half on the other.



Although not emphasized by the authors, ML and Bayesian estimation under partitioned models, estimated the phylogeny quite well, but ignoring the heterogeneous branches results in inconsistency.





Studies like this have led to the well-established use of partitioned models in phylogenetics.

B. Partitioned Models

Recall that the basic form of the likelihood function is:

$$P(D \mid \tau, M) = \prod_{i=1}^{s} P(D_i \mid \tau, M).$$

This is the likelihood under a *single* partition, where a single model applies to all the sites.

We can partition the data (say by gene) and use a separate model for each of *g* partitions:

$$P(D \mid \tau, M) = \prod_{i=1}^{s} \sum_{j=1}^{g} w_j P(D_i \mid \tau, M_j),$$

where w_j is he probability that site *i* is in partition *g* and the SSLs are weighted sums across partitions.

In conventional partitioned analyses, we assign sites *a priori* and all w_j's are either 0 or 1.

For a while, there were only Bayesian implementations of partitioned models, but now GARLi, RAxML, IQTree, PhyML, and PAUP* all support partitioned models.

C. Example from Lab Data Set

```
First Positions                Second Positions               Third
positions
Tree 1                         Tree 1                         Tree 1
------------------             ------------------             --------------
-ln L 1361.57931               -ln L 574.82758                -ln L 4014.14543
Base frequencies:              Base frequencies:              Base frequencies
A 0.297135                     A 0.210392                     A 0.435821
C 0.223702                     C 0.228875                     C 0.324911
G 0.235795                     G 0.143413                     G 0.030437
T 0.243369                     T 0.417320                     T 0.208831
Rate matrix R:                 Rate matrix R:                 Rate matrix R:
AC 1.15219                     AC 4.2179e+08                  AC 5.2290e-10
AG 3.38708                     AG 2.7458e+08                  AG 6506.44535
AT 0.78766                     AT 1.0059e+08                  AT 77.76142
CG 6.7725e-05                  CG 9.0472e+07                  CG 1.6437e-35
CT 8.99814                     CT 4.0779e+08                  CT 3272.68748
GT 1.00000                     GT 1.00000                     GT 1.00000
P_inv 0.289087                 P_inv 0.738767                 P_inv 0.00804384
Shape 0.336931                 Shape 0.734197                 Shape 0.464942
```

There are a couple things to note here:

First, look at the variation in base frequencies across codon positions. We can see that just by allowing each partition to have its own set of b.f. we'll see a huge improvement in fit.

Second, we can simply sum the ln*L*s for each partition to get the score for the partitioned model:

```
(-1361.57931) + (-574.82758) + (-4014.14543) = -5950.55228
```

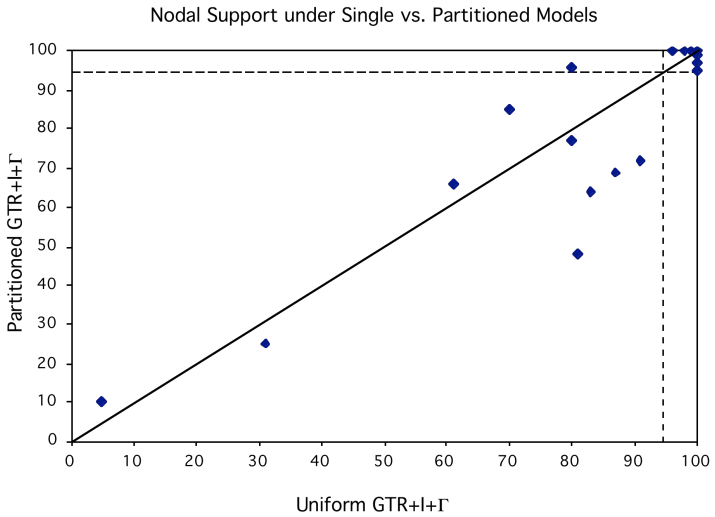
A single-partition GTR+I+Γ is shown.

```
Tree 1
```

```
------------------
-ln L 6467.45222
Base frequencies:
A 0.385640
C 0.329665
G 0.056255
T 0.228440
Rate matrix R:
AC 0.16037
AG 4.56668
AT 0.40618
CG 0.16011
CT 6.92991
GT 1.00000
P_inv 0.465785
Shape 0.489470
```

So, the unpartitioned ln*L* is 516.8999 units worse that the model partitioned by codons. We've gone from 10 to 30 parameters and compare the LRT statistic to $\chi^2_{[20]}$ and the *p*-value is <<0.001.

This certainly can influence phylogeny estimation: The figure below compares the support vales for all the nodes, and the results are mixed. For most of the well-supported nodes, they're well supported regardless of codon-based of single model. However, there are a few nodes for which this is not the case.



So, partitioning the likelihood model can make a difference for some nodes, and it's difficult to say which analysis is correct. We'll look at a more complex example in a minute.

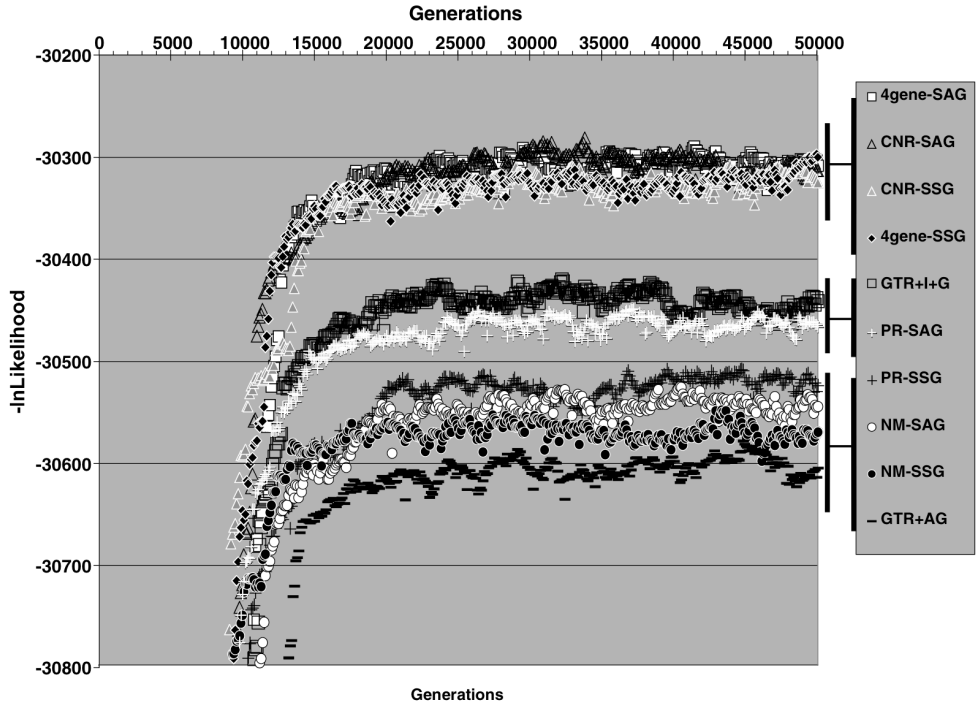
D. Selecting Partitioned Models - We can erect a huge number of alternative partitioning schemes, and just as is the case with selecting models, we need to select from among them.

Because most early implementations of partitioned models have been under a Bayesian framework, most early work on evaluating alternative partitioning schemes also focused on Bayesian model selection.

As can be seen above, partitioning leads to dramatic improvements in likelihood scores, but this improved fit may not improve phylogeny estimation.

We can use the same set of model selection approaches in we discussed earlier in the semester: LRT, AIC, BIC, DT.

Among the first to actually do this were Castoe et al. (2004. Syst. Biol. 53:448), who had a 4-gene data set.



They didn't conduct formal model selection but indicated that the 4 partitioned models all have about the same ln*L* so they chose the simplest of these.

McGuire et al. (2007) did a great job of assessing model-selection approaches to select among partitioning schemes.

Bayesian analyses No. partitions	No. parameters	$-HML_i$	AIC_c	BIC
1	11	94,809.5	189,641.1	192,408.3
2	24	92,880.2	185,808.7	188,657.3
4	48	91,659.9	183,417.0	186,416.3
5A	58	91,604.8	183,327.3	186,389.1
5B	60	91,509.5	183,140.8	186,215.1
6A	72	91,544.2	183,235.1	186,384.1
6B	70	91,478.3	183,099.1	186,235.7*
8	96	91,338.7	182,874.1	**186,172.3**
9	108	**91,296.0**	**182,814.0**	186,186.6
ML analyses No. partitions	No. parameters	$-L_i$	AIC_c	BIC
1	10	95,175.2	190,370.5	193,131.4
2	20	93,559.5	187,159.2	190,066.0
4	40	91,649.0	183,378.8	186,328.1
5A	48	91,602.7	183,302.6	186,301.9
5B	50	91,577.7	183,256.7	186,268.5*
6A	60	91,546.9	183,215.6	186,289.9
6B	58	91,542.4	183,202.5	**186,264.3**
8	80	91,468.7	183,100.7	186,299.5
9	90	**91,422.9**	**183,030.0**	186,290.9

As is usually the case, the AIC favored the most complex model evaluated, the BIC favored simpler partitioned models and the asterisks indicate the even simpler models favored by decision theory (DT).

E. Erecting Partitioning Schemes to Evaluate

One can envision a huge array of potential ways to partition data: by gene, by codon position, or even some combination of both.

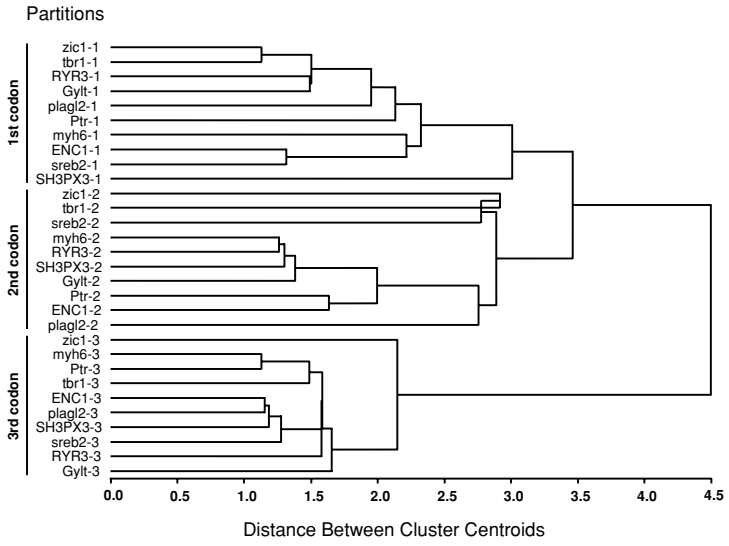
For, say 20 genes, there would be 60 partitions (3 codon positions per gene) and ~9.8×10^{59} possible partitioning schemes.

Many folks choose a few of these *a priori* to examine but Li et al. (2008. Syst. Biol. 57:519) proposed to cluster blocks hierarchically based on similarity of model parameters.

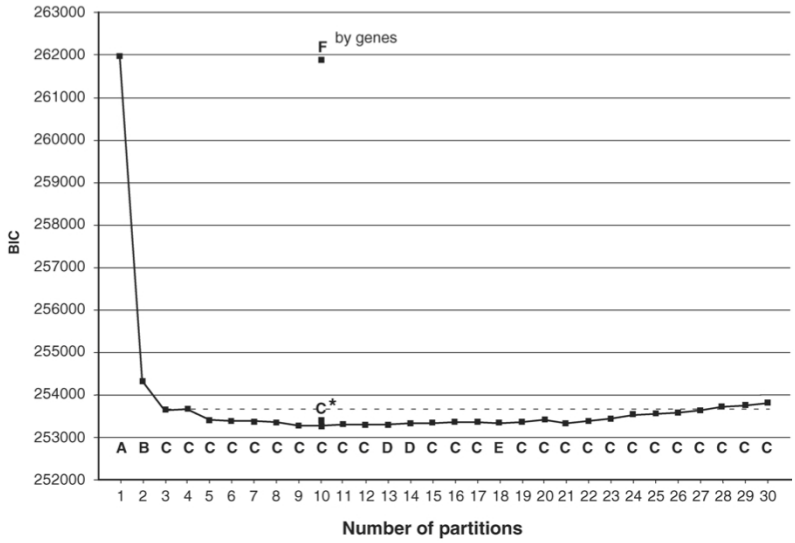
They had 10 genes, each split into codon positions, for a maximum of 30 partitions, and estimated parameters of GTR+Γ for each.

They then subjected these parameter estimates to centroid clustering to cluster partitions by similarity of parameter estimates.

This dendrogram represents that similarity and suggests a hierarchical array of partitioning schemes to examine.



This provides a logical way of generating 30 partitioning schemes. Note that when the blocks are lumped into 3 partitions, those correspond to codon positions.



BIC suggests use of 3-28 partition and partitioning just by gene doesn't seem very helpful.

Rob Lanfear (Lanfear et al. 2012. MB&E. 29:1695; Lanfear et al. 2014. BMCEvolBiol.; Lanfear et al. 2016) automates and enhances this approach with PartitionFinder.

PartitionFinder offers the option of doing a greedy search through partition schemes, and this is a very widely used approach (cited ~9000 times).

While it's an important advance, I'm convinced it's prone to over-partitioning; I've seen it erect partitions with few variable sites (say 15) but apply a GTR+I+Γ (with its 10 parameters) model to it.

It also (implicitly) promotes this notion that there is a "correct" partitioning scheme.

F. Mixture Models.

Remember that for partitioned models, the likelihood function is this:

$$P(D \mid \tau, M) = \prod_{i=1}^{s} \sum_{j=1}^{g} w_j P(D_i \mid \tau, M_j)$$

where the w_j's are the probability that site i evolved under model j (i.e., belong to partition j), and, when we assign sites to partitions, these are either 0 or 1.

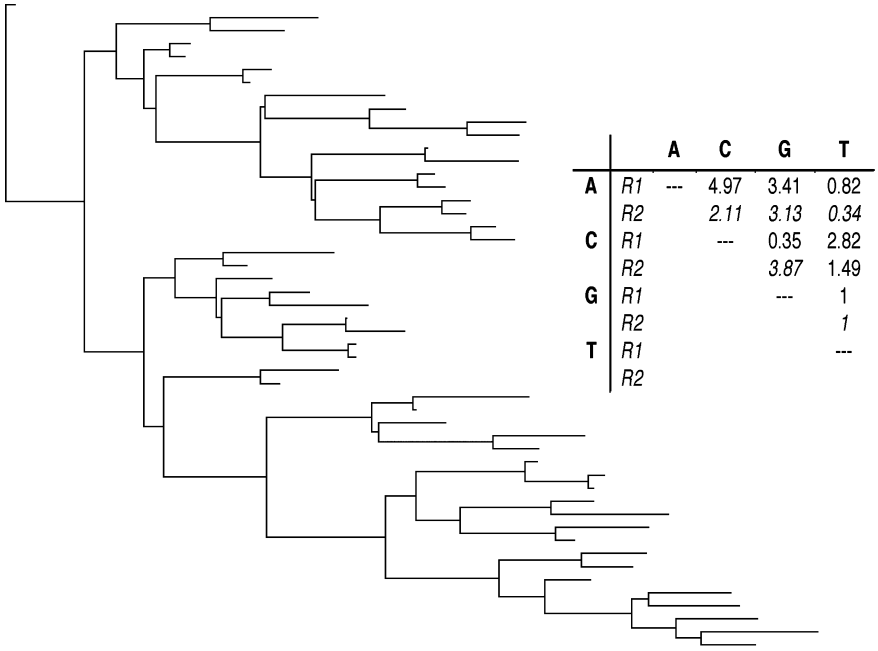
We've already introduced mixture models when we discussed Γ-distributed rates models of A-SRV and the approach is similar here.

Pagel & Mead (2004. Syst. Biol., 53:571) first applied mixture-models to the **Q**-matrix in phylogenetics.

Just as in Γ-distributed rates, this mixture model relaxes the requirement of assigning sites to partitions, and the w_j's are treated as random variables that are estimated for each site.

So now we have multiple **Q**-matrices, all of which are unrestricted (that is, each model is a full GTR with six rate transformation rate parameters).

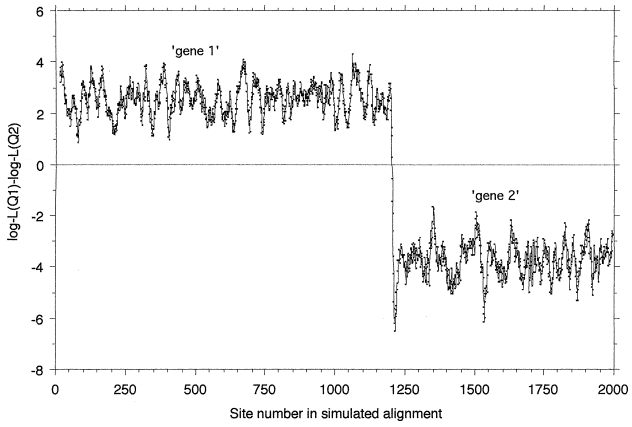
They used simulation to test the ability of their mixture model to detect partitions. They simulated data from two different models.



		A	C	G	T
A	*R1*	---	4.97	3.41	0.82
	R2		*2.11*	*3.13*	*0.34*
C	*R1*		---	0.35	2.82
	R2			*3.87*	1.49
G	*R1*			---	1
	R2				*1*
T	*R1*				---
	R2				

They simulated 1200 bp with the first R-matrix (assuming equal b.f.) and 800 bp with the second.

When they ran their MCMC under a GTR-2Q mixture model, they were able to identify precisely the point in the concatenation where the generating model changed.



So, these results (and others from analyses of real data) are pretty promising and suggest that use of mixture models may result in better phylogeny estimation.

One question we've not addressed is how to determine the degree of the mixture; that is how many distinct GTR models should we include.

There are a couple papers that have applied rjMCMC to mixture models an let the degree of the mixture be estimated directly from the data.

Venditti et al. (2008. Syst. Biol., 57:286) was the first to do this, and they maintained the requirement of a single set of b.f. and that each R-matrix be unconstrained.

Evans and Sullivan (2012. Syst. Biol., 61:12-21) have generalized this so that each model can have its own b.f. and can take on any of the 203 restrictions of the R-matrix. This is probably over parameterized for most data sets, but is may be required when we have lots of phylogenomic data.

Caveat – Use of mixture models and their special cases of partitioned models and assume all partitions evolved on the same tree; that is, this ignores coalescent stochasticity and hybridization.